**Research Proposal: Part-Based Modular GCN + Transformer Fusion for 3D Human Pose Estimation**

**Title**

**Anatomically Modular Graph Convolutional and Transformer-Based Architecture for 3D Human Pose Estimation**

**Abstract**

We propose a novel deep learning architecture for accurate frame-wise 3D human pose estimation from 2D joint keypoints. The model partitions the human skeleton into anatomically meaningful subgraphs (e.g., arms, legs, torso), which are individually processed by independent Graph Convolutional Networks (GCNs). These part-level embeddings are then fused by a global Transformer encoder that captures inter-part dependencies. Our pipeline is designed for the MPI-INF-3DHP dataset and achieves enhanced accuracy, modularity, and generalizability compared to baseline GCN models.

**Motivation**

Traditional GCN-based approaches model the full human skeleton as a single graph, which overlooks localized structural patterns unique to different body regions. Furthermore, Transformer-based methods often ignore local part-specific constraints. Our proposed method balances both local and global structural modeling by separating the skeleton into modular parts, learning their representations independently, and then fusing them for holistic pose reasoning.

**Dataset**

**MPI-INF-3DHP** (28 joints, 3D ground truth) is used for training and evaluation. Data preprocessing follows standard normalization and zero-padding for the input keypoints.

**Architecture Overview**

**1. Input**

* Shape: (B, 28, 3) representing x, y, z coordinates for 28 joints.

**2. Graph Partitioning**

* **Upper Body:** Joints 0-1, 2-7, 12-15
* **Lower Body:** Joints 8-11, 16-19, 20-23
* **Torso:** Joints 1, 8, 9, 12, 13
* Each part is treated as an independent subgraph.

**3. Part-Level GCN Modules**

Each subgraph is processed by a dedicated 2-layer Sparse GCN:

* Layer 1: GCNConv(in\_channels, hidden\_dim)
* Layer 2: GCNConv(hidden\_dim, hidden\_dim)
* Output: (B, N\_part, hidden\_dim)

**4. Concatenation & Positional Encoding**

* Concatenate all outputs: (B, 28, hidden\_dim)
* Add sinusoidal or learned positional encodings.

**5. Transformer Encoder**

* Operates across all 28 joints with multi-head attention.
* Learns inter-part spatial dependencies.
* Output: (B, 28, hidden\_dim)

**6. MLP Head**

* A shared MLP head maps each joint embedding to its final 3D coordinate: (B, 28, 3)

**Loss Functions**

* **MPJPE** (Mean Per Joint Position Error)
* **Bone Length Regularization** to maintain anatomical realism
* Optional: Joint symmetry or angular constraint loss

**Novel Contributions**

1. Anatomical decomposition of human pose into GCN-processed subgraphs
2. Transformer-based fusion for global context and inter-limb coordination
3. Modular architecture for better interpretability and generalization
4. Compatible with real-time inference and visualization pipelines

**Future Work**

* Extend to video-based inference using a temporal Transformer
* Export model to ONNX for real-time deployment
* Add joint uncertainty modeling via confidence-based masking

**Proposed Repository Structure**

├── model.py # GCN + Transformer architecture

├── train\_pose\_model\_parts.py # Training script

├── predict\_image\_parts.py # Inference on image

├── utils/ # Edge indices, normalization, losses

├── data/ # MPI-INF dataset & normalization stats

├── README.md # Documentation and usage

**Citation (Tentative)**

Vedansh Tembhre, 2025. "Anatomically Modular Graph and Transformer-Based Architecture for 3D Human Pose Estimation."

**Contact**

*Vedansh Tembhre*  
[vedanshtembhre@my.unt.edu](mailto:vedanshtembhre@my.unt.edu) | [https://vedanshtembhreportfolio.com](https://vedanshtembhreportfolio.com/)